Vision for Collaboratories and the DOE Science Grid

William E. Johnston
National Energy Research Scientific Computing Division
Lawrence Berkeley National Laboratory

The Vision for a DOE Science Grid

Large-scale science and engineering is typically done through the interaction of

- **people,**
- **heterogeneous computing resources,**
- **multiple information and storage systems, and**
- **instruments,**

all of which are geographically and organizationally dispersed.

The overall motivation for "Grids" ([2],[3]) is to *enable the routine interactions* of these resources to facilitate this type of large-scale science and engineering.

Two Sets of Goals

The overall goal is to facilitate the establishment of a DOE Science Grid **("DSG") that ultimately incorporates production resources and involves most, if not all, of the DoE Labs and their partners. (See "Vision and Strategy for a DOE Science Grid" [2])**

A second goal is to use the Science Grid framework to support the R&D agenda and construction of Collaboratory Grids: Grid infrastructure as seen from the point of view of scientific collaboration and distributed, scientific/laboratory experiments.

Applications

Several types of science and engineering scenarios are motivating **the development and deployment of Grids at DOE and NASA:**

- ***Large-scale, multi-institutional engineering design and multi-disciplinary science that require locating and co-scheduling many resources*** **– e.g., design of next generation diesel engines, next generation space shuttle, etc.**

- ***Scientific data analysis and computational modeling with a world-wide scope of participants*** **– e.g. High Energy Physics data analysis and climate modeling**
- ***Real-time data analysis for on-line instruments*****, especially those that are unique national resources – e.g. LBNL's and ANL's synchrotron light sources, PNNL's gigahertz NMR machines, etc.**
- ***Coupling of laboratory instrument experiments*** **and computational models to support, e.g., experiment and computational steering**

- ***Generation, management and use of very large, complex data archives*** **that are shared across an entire community – e.g. DOE's human genome data and NASA's EOS data**
- ***Collaborative, interactive*** **analysis and visualization of massive datasets – e.g. DOE's Combustion Corridor project.**

Addressing the requirements of these application classes in a general way with common Grid infrastructure deployed across the DOE Labs and collaborating institutions will enable many different applications to routinely use widely distributed resources**.**

Examples of Grid-like Systems

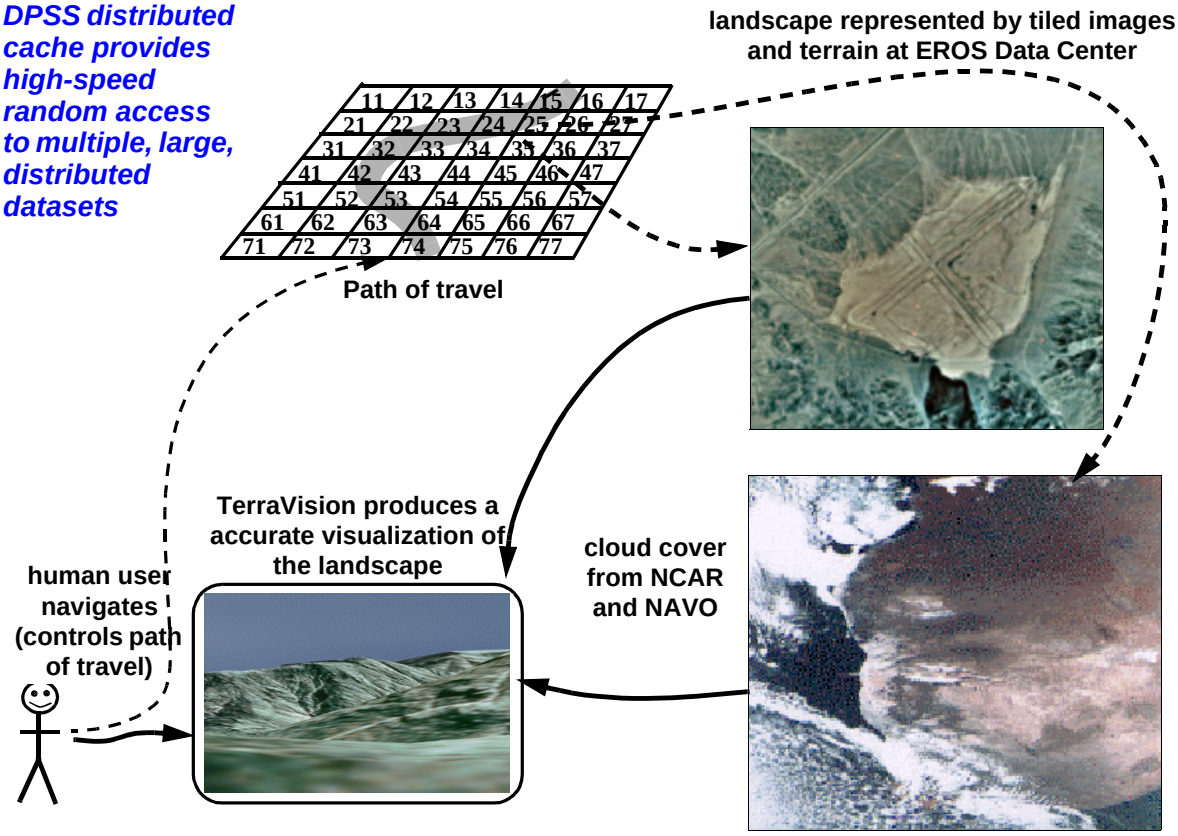
¨ ***High data-rate, widely distributed data management*** **(federated access for archived satellite and aerial imagery, digital terrain data, and atmospheric data in the MAGIC Gigabit Testbed [10], [11])**

- **on demand, real-time interactive exploration of an operational environment supporting, e.g., military operations and community emergency services**
- **aggregation of multiple, widely distributed, multi-discipline data sets**

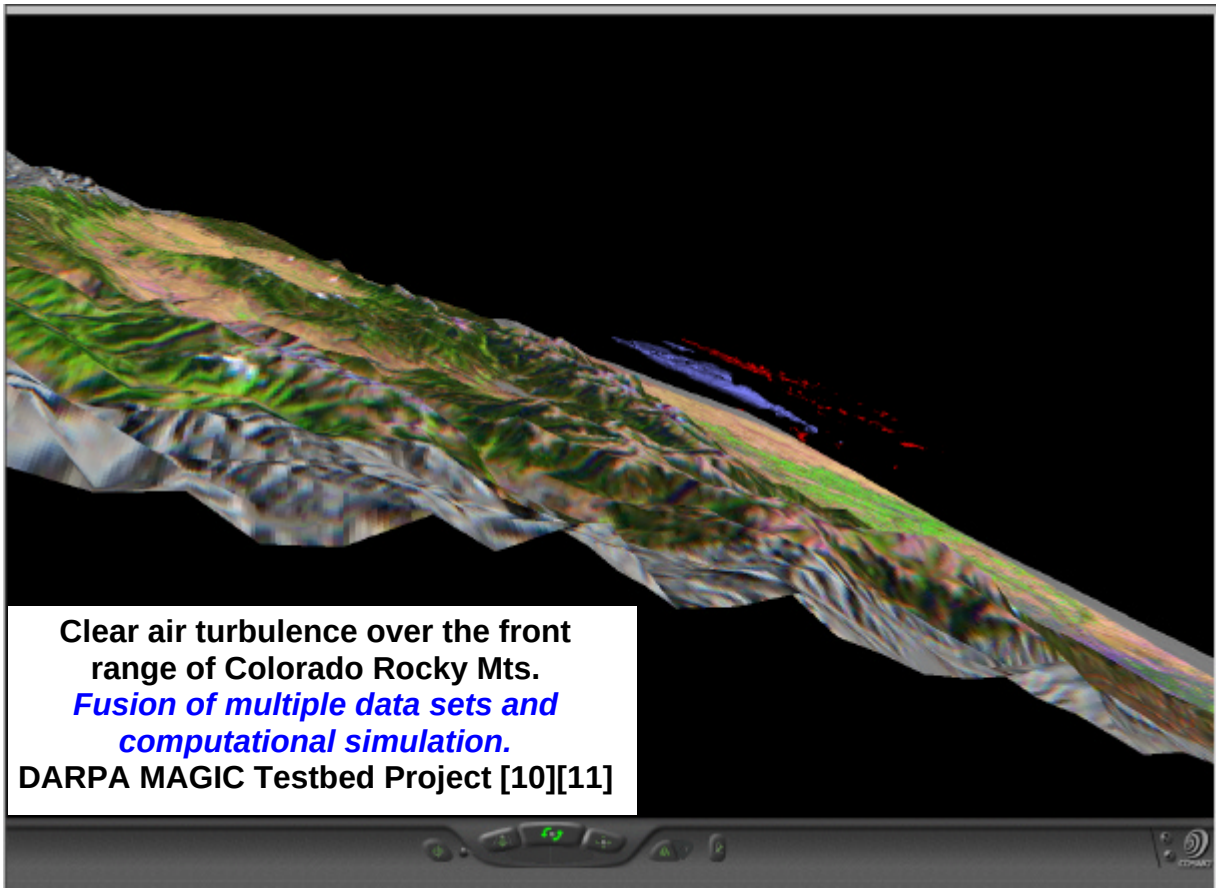
MAGIC Testbed

- **on-line, real-time access to multiple environmental data sets that are (and always will be) maintained by domain experts at their own sites.**
- **DARPA MAGIC testbed consortium (see www.magic.net) developed distributed services, data and visualization from EROS Data Center, NCAR, NAVO, SRI**
- ***Similar characteristics to DOE applications like the Combustion Corridor [6], the Physics Particle Data Grid [7], the Earth Sciences Data Grid, etc.***

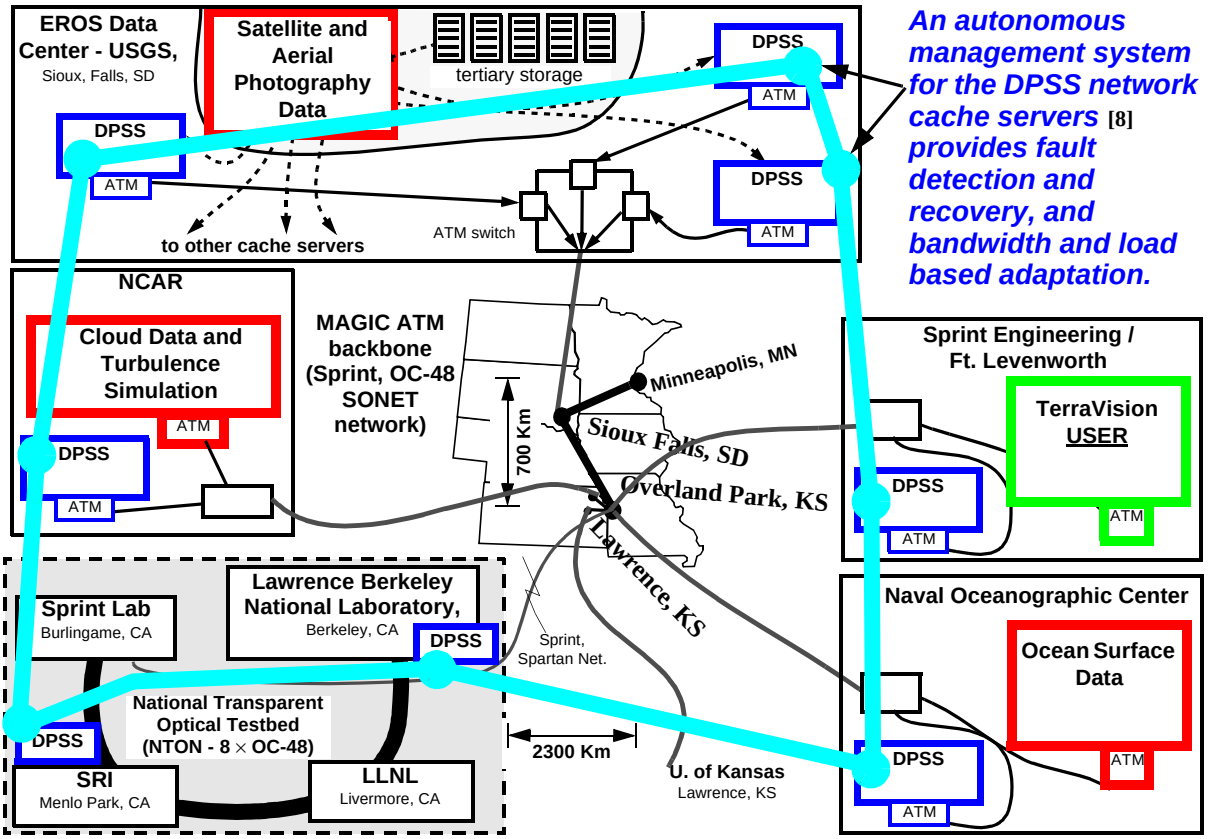
DPSS distributed cache provides high-speed random access to multiple, large, distributed datasets



TerraVision Provides Real-time Visualization of Aggregated Data



Clear air turbulence over the front range of Colorado Rocky Mts.
Fusion of multiple data sets and computational simulation.
DARPA MAGIC Testbed Project [10][11]



The MAGIC Testbed Distributed Application Environment

Examples of Grid-like Systems

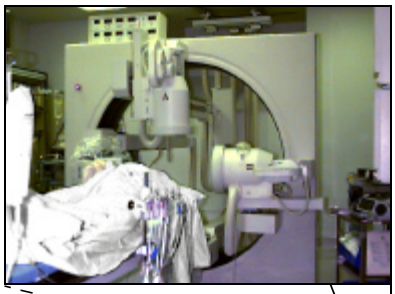
- **On-line medical imaging system (*real-time digital libraries for on-line, high data-rate instruments* [9])**
  - **on-line, real-time, high data-rate medical instrument with remote users**
  - **distributed data analysis and automatic data cataloguing and archiving**
  - **strict authorization and access control**
  - **optical WDM metropolitan area network (NTON)**
  - ***Similar characteristics to DOE projects like Clipper [13]***

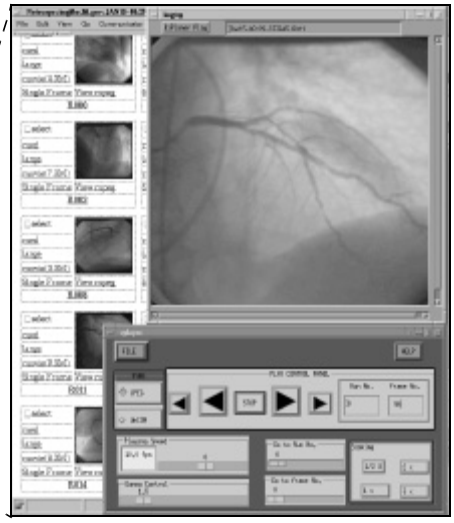
WALDO real-time digital library system and DPSS distributed cache [9] for data cataloguing and storage

Tertiary Storage

Compute servers for data analysis and transformation

Kaiser San Francisco Hospital Cardiac Catheterization Lab (X-ray video imaging system, ≈ 130 mbit/s, 50% duty cycle 8-10 hr/day)





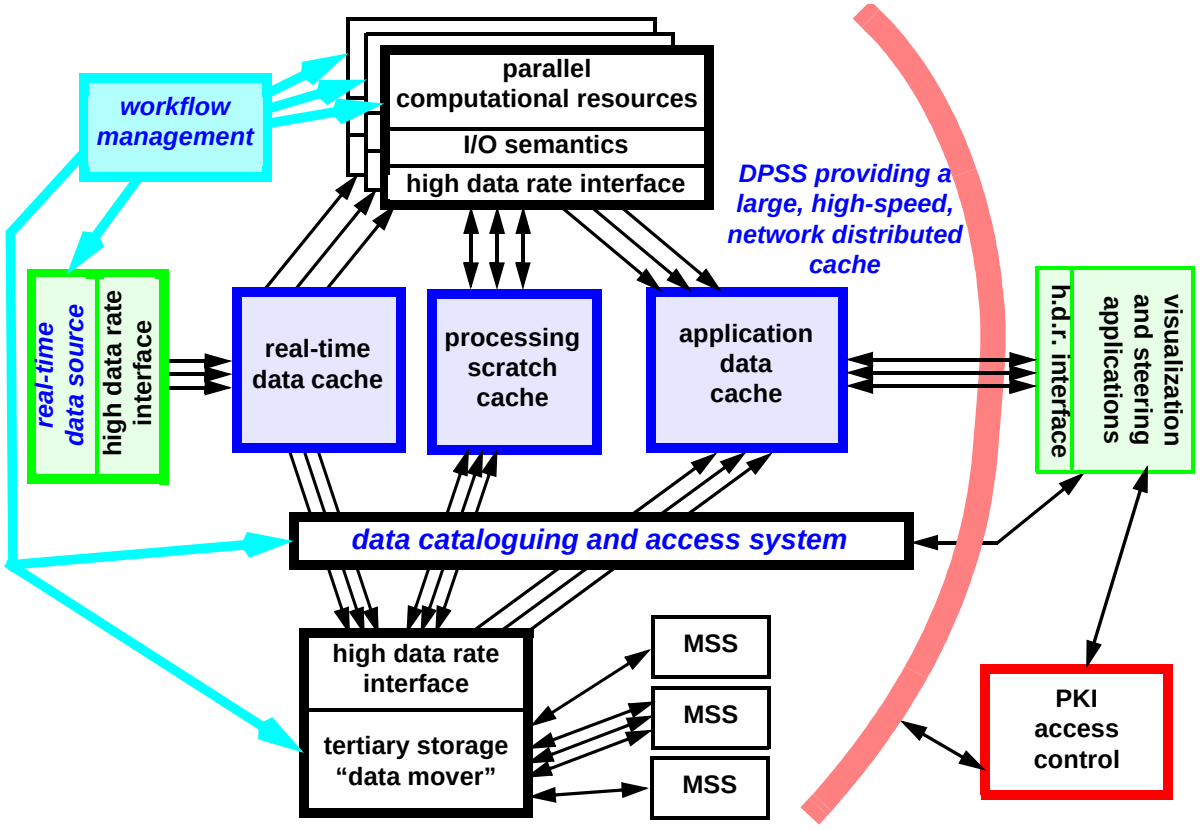
The PSE: Automatically generated user interfaces providing indexed access to the large data objects (the X-ray video) and to various derived data.

Kaiser Oakland Hospital (physicians and databases)

Kaiser Division of Research

Lawrence Berkeley National Laboratory and Kaiser Permanente Health Care
On-line Health Care Imaging Experiment

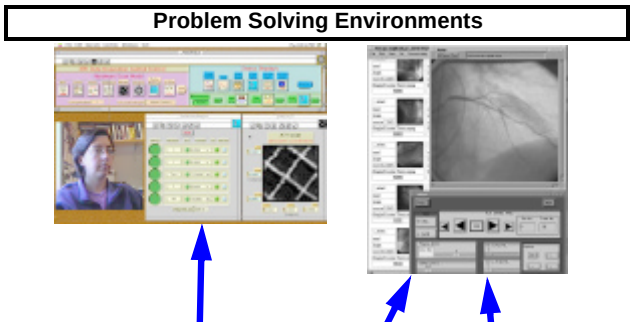
NTON network testbed



A High Volume, High Data Rate, Data Analysis Architecture

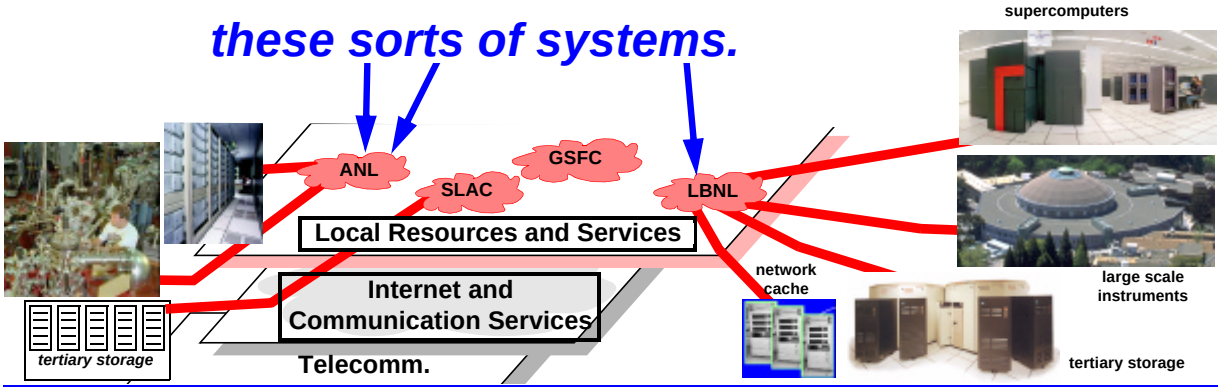
These large-scale science and engineering problems involve many types of applications and data sources that are accessed and shared across many institutions. This implies:

- numerous interconnected servers providing computational simulation and analysis, data access, and functional access to instruments, in semi-open agency/research networks (e.g., ESNet, NREN, Internet-2, etc.)
- many simultaneous collaborators, e.g. at DOE Labs, NASA Centers, other Federal labs, industrial partners, and universities
- many stakeholders and diverse assets .



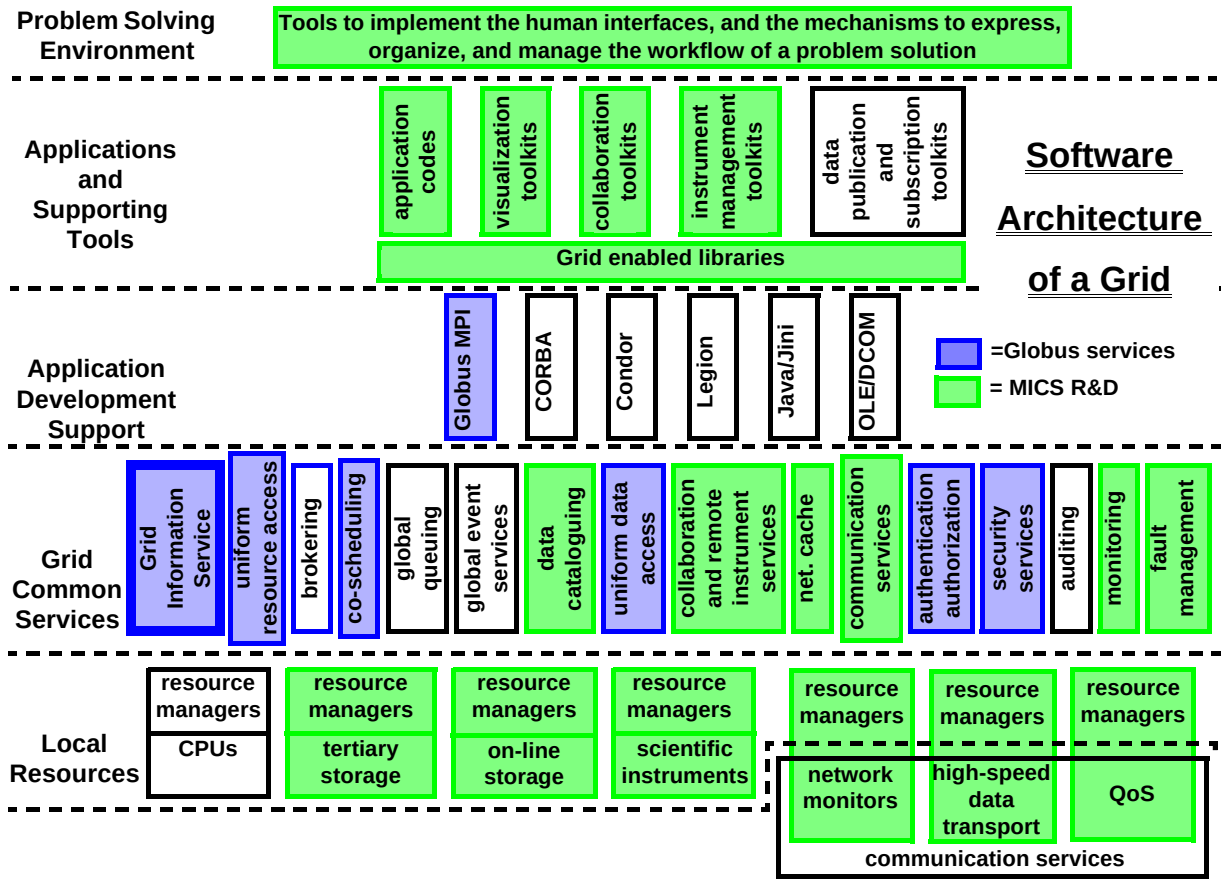
The Evolution of Distributed Environments

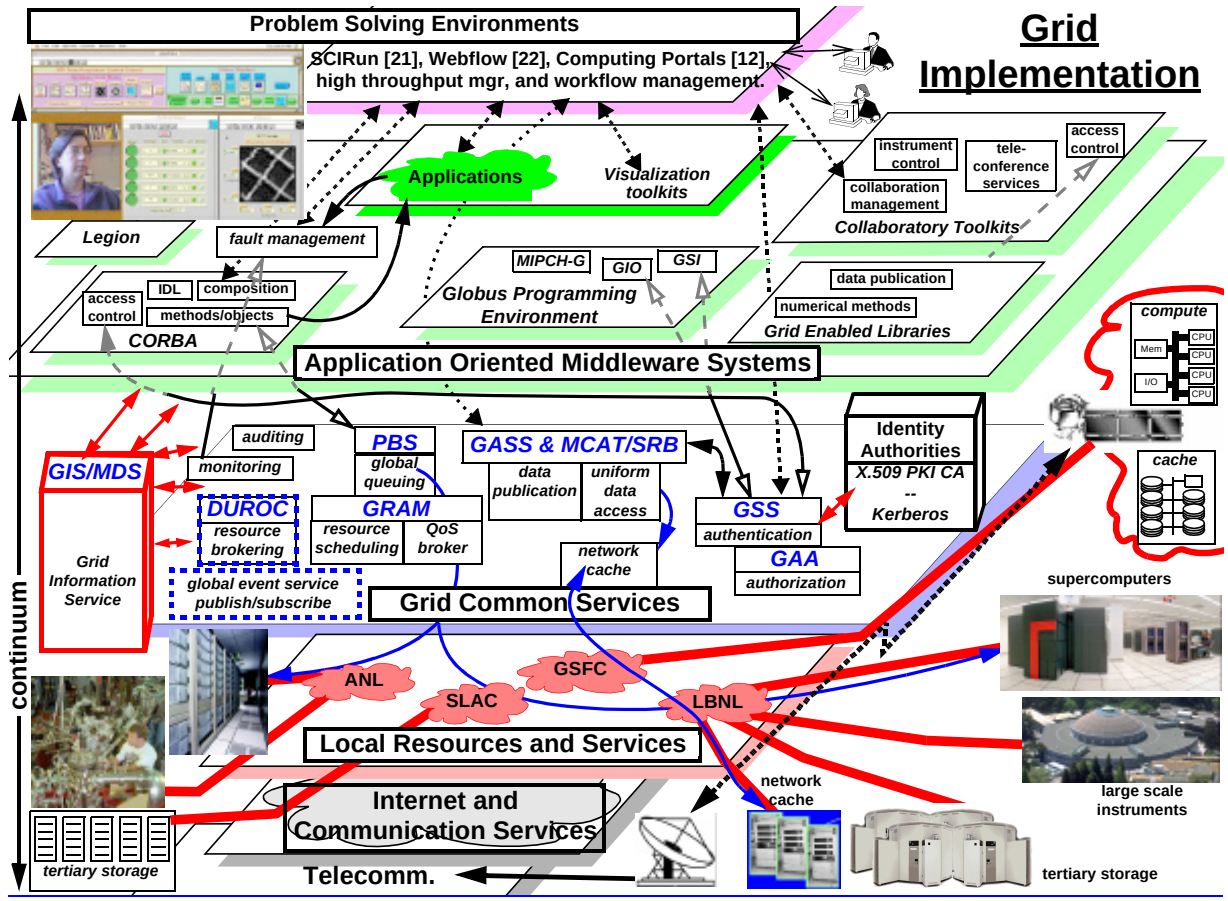
The DOE Science Grid should provide a shared, and when possible, uniform view of resources in order to facilitate these sorts of systems.



Grids

- ***Grids will provide a rich set of architecturally consistent services*** **for constructing, using, and managing the types of diverse, widely distributed environments described in the preceding examples**
- ***Grids will provide uniform access and co-scheduling for computing, data, communications, and collaboratory resources.***
- ***Grids are built through collaborative efforts*****, and at the same time facilitate collaboration.**





Grids

- ***To be useful for applications, the DOE Science Grid must also provide operational and user support*** **and persistent infrastructure.**
- ***For such an open environment to be feasible, security must be a design goal from the start*****, and must address authentication, authorization, and infrastructure assurance**

Collaboratory Grids

Science Grid capabilities that support collaboratories and collaboration

- **Collaboratory services**
- **Data and information management**
- **Collaborative visualization and data analysis**
- **Collaborative PSEs for laboratory experiments**
- **Communication services**
- **Authorization and access control**
- **Infrastructure management**

Collaboratory services

Light-weight collaboration environment

- **Issues**
  - **much of the time collaborators do not need synchronous, audio/video interactions**
  - **video conferencing support has a high cost, and, even so, by itself is not enough**
  - **need tools that provide a continuum of interaction capabilities**

Collaboratory services

- **Goals**
  - **a “light-weight” environment that promotes experimentation with new models of interaction**
    - **shared operational tasks (“the global control room”)**
    - **new human <-> machine interfaces**
      - **policy based navigation within persistent spaces**
      - **direct participation by processes in shared collaboration spaces**

Collaboratory services (Goals, cont.)

- **flexible and secure access control for collaboration environment**
- **easily run from anywhere, on any system, at any time**
- **persistent, interactive space that supports a range of contact mechanisms**

- **Approach**
  - **leverage off of Grid infrastructure**
    - entity registration
    - synchronous and asynchronous contact
  - **document and application sharing**
  - **integrated video conferencing in the light-weight environment**
  - **accessible via a Web browser**
  - **use existing tools and standards**

Data and information management

Data location management

- **Issues**
  - **managing and using terabytes of data that have be shared among hundreds of users world wide will be dominated by managing where data originates, where it is analyzed, and where it is cached**

- **Goals**
  - **locate data and schedule data movement from tertiary storage**
  - **minimize the application access time**
  - **maximize sharing by multiple users from local caches**
  - **coordinate cache space reservation and file movement with network QoS**

- **Approach**
  - **Storage Resource Managers (“SRMs”) that manage storage resources, such as HPSS tertiary storage systems, DPSS network caches, shared local disk caches**
  - **provide a reservation capability**
  - **manage request queues**
  - **monitor transfer errors**
  - **reschedule transfer in case of system failures**
  - **use Science Grid services such as replicate catalogues, security services, communication services (e.g. QoS)**

- **build on LBL's current and previous work in this area:**
  - **Storage Access Coordination System (STACS - [15]) has been adapted to the Grid**
  - **SDSC's SRB/MCAT has been interfaced to the STACS HRM (HPSS Resource Manager)**
  - **SRB access has been extended to include pre-staging and time estimation from HRM**
  - **a new data "Request Manager" has been interfaced to Globus replica catalog and file transfer services**

Collaborative visualization and data analysis

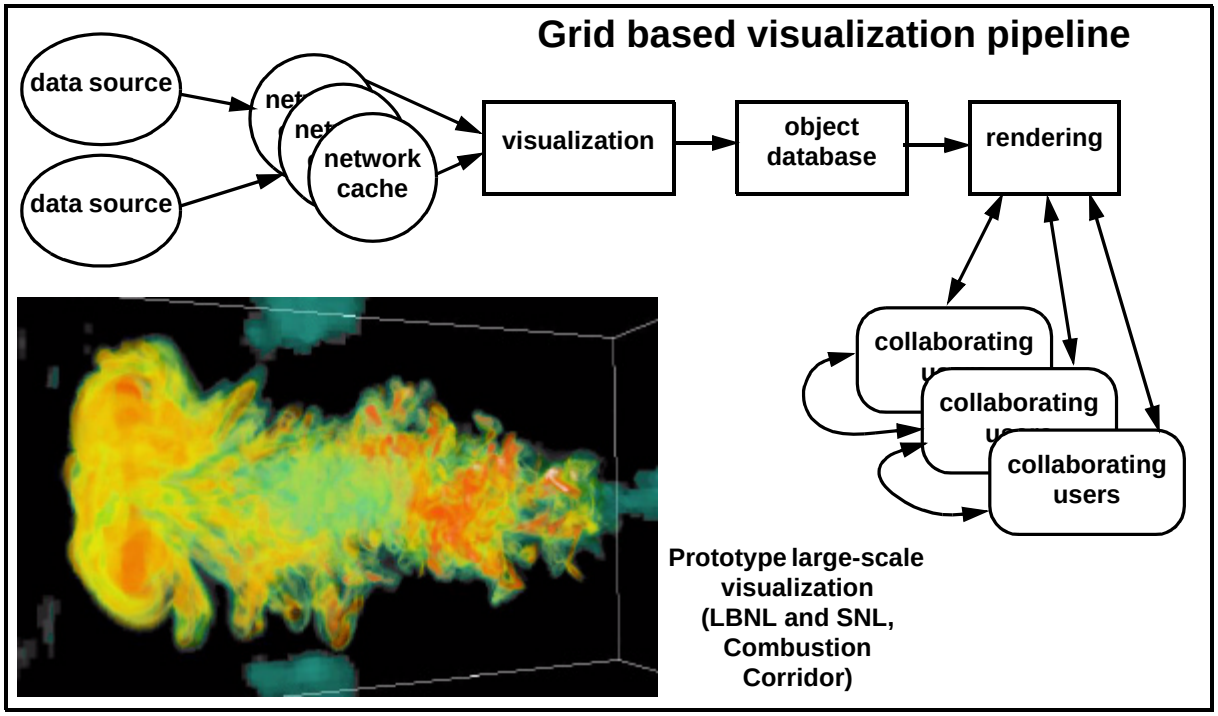
¨ **Goals**

- **enable the routine distributed visualization, analysis, and collaboration on terabyte size datasets**

- **Issues**
  - **access to data at sufficiently high rates to permit interactive manipulation**
  - **coupling of computing resources to data sources**
  - **temporary storage of data on high-speed, random access devices**
  - **scheduling all resources, including specialized graphics hardware**

- **Approach**
  - **use Grid data storage management services to locate, request, and migrate archived data to network cache storage (DPSS – [8])**
  - **use reservable network caches to provide high-speed, random access**
  - **use Grid co-scheduling services to simultaneously reserve cache storage, network bandwidth, computing capacity, and graphics rendering systems**

- use combustion data and high-speed networks, compute servers, and render servers as testbed (LBNL+SNL)



Collaborative PSEs for laboratory experiments

Laboratory workflow representation and management for instruments, analysis, and simulation

¨ **Goals**

- flexible, expressive, and preservable representation of generalized experiment protocols
- coupling unique experimental facilities with computational platforms to achieve novel functionality

Collaborative PSEs for laboratory experiments (Goals, cont.)

- **feature based characterization of vast amounts of data for efficient visualization, comparative analysis, and data mining**

¨ **Issues**

- **rule-based workflow management**
- **how computational components are to be coupled with the experimental facilities**
- **what image/data features need to be characterized?**

Collaborative PSEs for laboratory experiments (Goals, cont.)

¨ **Approach**

- **Leverage scientific experiments in material science and cell biology for abstraction and design of reusable software components (in-situ experiments, cell signaling, multi-cellular signaling, and chemosensitivity diagnostics)**
- **build on previous work in visual algorithms to meet feature based representation of images and video streams (LBNL and SNL)**
- **use Science Grid for uniform access to computing and storage resources, and for security services**

- **leverage XML work in workflow representation (Wf-XML) and PMML (Predictive Markup Modeling Language) for exchange of declarative knowledge, rules, and models in a heterogeneous environment (LBNL and ORNL)**
- **build on the DeepView CPSE ([16]) by using an inference engine to address new types of science experiments (LBNL, SNL, and ORNL)**

Communication services

Secure and reliable group communication

¨ **Issues**

- **collaborators are a peer group with multiple senders and current IETF efforts focussing on single sender groups**
- **need to be able to control who will be allowed into a group**
- **collaborators will be spread across the Internet**
- **different applications have different reliability and message ordering requirements**

Communication services (Issues, cont.)

- difficulty of doing membership and latency of message delivery for reliable multicast protocols supporting multi-sender groups increases with the size of the membership of the group

¨ Goals

- simplify collaboratory group communication
- toolkit for secure, easy to use, flexible reliable multicast
- support for single-sender and multi-sender groups
- protocol that scales to large groups spread across the Internet

Communication services

¨ Approach

- split membership into a sender and a receiver group
- allow receivers to choose their own desired level of reliable delivery and message ordering
- secure group membership and secure intra-group messaging via the Cliques group key management protocol (LBNL, ISI, and UC Irvine)
- integrate Akenti for authorization of group members

- initial prototype implementation of protocol in Java with C++/C introduced as needed to achieve performance and portability goals
- investigate flow and congestion control issues

Authorization and access control

Policy based authorization in multi-stakeholder, multi-user distributed environments

¨ **Issues**

- distributed management – because the principals and resources are dispersed organizationally
- distributed access control – because the resources and users are dispersed geographically
- the multiplicity of stakeholders in major resources must all be able to ensure that their use conditions are met

Authorization and access control (Issues, cont.)

- **qualified users must be able to “transparently” gain access**
- **unqualified entities must be strongly prevented access**

¨ **Goals**

- **dynamic and easily used mechanisms for generation, maintenance, and distribution of the access control information**
- **strong assurances that use-conditions are met**
- **a policy-neutral mechanism**

Authorization and access control (Goals, cont.)

¨ **Approach**

Akenti is a policy driven authorization system being developed for the collaboratory environment [16].

- **stakeholders are associated with resources by trusted third parties (currently in configuration files associated with the resource)**
- **all other “trust” is explicit**

- Akenti is basically a data driven certificate analyzer
  - user identity and resource identity are presented
  - stakeholders are identified
  - use-conditions are collected and verified
  - required attributed are located and verified
- result of a successful authorization is packaged as a “capability” and passed to the resource access control gateway (which may enforce run-time conditions, e.g. role based controls)
- integrate with the IETF, Generic Authorization and Access control API [16]

Infrastructure management

Services to build fault tolerant systems, and do performance monitoring, in widely distributed and organizationally heterogeneous environments

¨ **Issues**

- servers running on systems where the user has no administrative influence must be managed autonomously
- a large collection of servers that must operate in concert to accomplish a task, and that run on many, widely distributed systems must be managed autonomously

- **Goals**
  - **reliably functioning systems that can be built on-demand from a set of resources that are brokered based on general user characteristics and resource applicability**
  - **adaptive and customizable monitoring services**
- **Approach**
  - **agent based server monitors that can monitor and correct the state of other servers**
  - **agent based performance monitors that have adaptive and/or dynamically modifiable behavior**
  - **mechanisms for agent reliability**

Collaboratories Grid Vision

- ***Routine collaboration among DOE Labs and their institutional partners through ready access to collaboration tools and remote instrument operation***
- ***New approaches to laboratory science made possible by routine, location independent access to large-scale computing and storage systems that can provide for real-time analysis of experiment data and feedback based experiment control***

References and Acronyms

[1] Globus is a middleware system that provides a suite of services designed to support high performance, distributed applications. Globus provides:

- Resource Management: Components that provide standardized interfaces to various local resource management systems (GRAM) manage allocation of collections of resources (DUROC). All Globus resource management tools are tied together by a uniform resource specification language (RSL).
- Remote Access: Components that enable remote access to files (GASS and RIO) and executables (GEM).
- Security: Support for single sign-on, authentication, and authorization within the Globus system (GSI) and (experimentally) authorization (GAA).
- Fault Detection: Basic support for building fault detection and recovery into Globus applications.
- Information Infrastructure: Global access to information about the state and configuration of system components of an application (MDS).
- Grid programming services: Support writing parallel-distributed programs (MPICH-G), monitoring (HBM), etc.

www.globus.org provides full information about the Globus system.

[2] *The Grid: Blueprint for a New Computing Infrastructure*, edited by Ian Foster and Carl Kesselman. Morgan Kaufmann, Pub. August 1998. ISBN 1-55860-475-8. http://www.mkp.com/books_catalog/1-55860-475-8.asp

[3] "Grids as Production Computing Environments: The Engineering Aspects of NASA's Information Power Grid," William E. Johnston, Dennis Gannon, and Bill Nitzberg. Eighth IEEE International Symposium on High Performance Distributed Computing, Aug. 3-6, 1999, Redondo Beach, California. (Available at http://www.nas.nasa.gov/~wej/IPG)

[4] "Vision and Strategy for a DOE Science Grid" - http://www.itg.lbl.gov/~wej/Grids

[5] See www.nas.nasa.gov/IPG for project information and pointers.

[6] See http://www-itg.lbl.gov/NGI/ for project information and pointers.

[7] The Particle Physics Data Grid has two long-term objectives. Firstly: the delivery of an infrastructure for very widely distributed analysis of particle physics data at multi-petabyte scales by hundreds to thousands of physicists. Secondly: the acceleration of the development of network and middleware infrastructure aimed broadly at data-intensive collaborative science. http://www.cacr.caltech.edu/ppdg/

[8] Tierney, B. Lee, J., Crowley, B., Holding, M., Hylton, J., Drake, F., "A Network-Aware Distributed Storage Cache for Data Intensive Environments", Proceeding of IEEE High Performance Distributed Computing conference (HPDC-8), August 1999.

[9] "Real-Time Generation and Cataloguing of Large Data-Objects in Widely Distributed Environments," W. Johnston, Jin G., C. Larsen, J. Lee, G. Hoo, M. Thompson, and B. Tierney (LBNL) and J. Terdiman (Kaiser Permanente Division of Research). Invited paper, International Journal of Digital Libraries - Special Issue on "Digital Libraries in Medicine". May, 1998. http://www-itg.lbl.gov/WALDO/

[10] MAGIC: "The MAGIC Gigabit Network." See: http://www.magic.net

[11] TerraVision-2: VRML based data fusion and browsing - www.ai.sri.com/TerraVision

[12] A collaborative effort to enable desktop access to remote resources including, supercomputers, network of workstations, smart instruments, data resources, and more - computingportals.org

[13] The Clipper Project: Computational Grids providing middleware that supports applications requiring configurable, distributed, high-performance computing and data resources. See http://www-itg.lbl.gov/~johnston/Clipper

[14] The Grid Forum (www.gridforum.org) is an informal consortium of institutions and individuals working on wide area computing and computational Grids.

[15] "New Capabilities in the HENP Grand Challenge Storage Access System and its Application at RHIC" http://rncus1.lbl.gov/GC/docs/chep292lp1.doc
"STACS is ... responsible for determining, for each query request, which events and files need to be accessed, to determine the order of files to be cached dynamically so as to maximize their sharing by queries, to request the caching of files from

HPSS in tape optimized order, and to determine dynamically which files to keep in the disk cache to maximize file usage."

[16] "DeepView: A Collaborative Framework for Distributed Microscopy." IEEE Conf. on High Performance Computing and Networking, Nov. 1998. See http://vision.lbl.gov/ (projects -> collaborative computing)

[17] **Akenti**: "**Certificate-based Access Control for Widely Distributed Resources**," Mary Thompson, William Johnston, Srilekha Mudumbai, Gary Hoo, Keith Jackson, Usenix Security Symposium '99. Mar. 16, 1999. (See http://www-itg.lbl.gov/Akenti)

[18] GAA: "**Generic Authorization and Access control API**" (GAA API). IETF Draft. http://ghost.isi.edu/info/gss_api.html)

[19] Storage Resource Broker (SRB) provides uniform access mechanism to diverse and distributed data sources. http://www.sdsc.edu/MDAS/

[20] Condor is a High Throughput Computing environment that can manage very large collections of distributively owned workstations. http://www.cs.wisc.edu/condor/

[21] SCIRun is a scientific programming environment that allows the interactive construction, debugging and steering of large-scale scientific computations. http://www.cs.utah.edu/~sci/software/

[22] WebFlow - A prototype visual graph based dataflow environment, WebFlow, uses the mesh of Java Web Servers as a control and coordination middleware, WebVM. See http://iwt.npac.syr.edu/projects/webflow/index.htm